CLAIMS

What is claimed is:

1. A computer-implemented method of identifying table data in a document comprising the steps of:

a) receiving a page description language representation of the document for providing a list of words in the document and position information for the words; and

b) automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature.

2. The method of Claim 1 wherein the step of automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature further comprises:

b1) dividing the document into one or more pages;

b2) dividing each page into a plurality of lines;

b3) for each line, clustering the words of the line into one or more word clusters;

b4) automatically identifying table data in the document based on the number of word clusters for each line and the alignment of the word clusters between lines.

3. The method of Claim 2 wherein the step of automatically identifying table data in the document based on the number of word clusters for each line and the alignment of the word clusters between lines further comprises:

b4_1) using the word clusters to generate column position information; and

b4_2) updating the column position information by performing a union operation between the column position information of the previous line and the column position information of the current line.

4. The method of Claim 1 wherein said step of automatically identifying table data in the document based on the page description of the document and at least one table identifying feature comprises:

b1) automatically determining a table bounding box for each table in the document;

b2) expanding each table bounding box based on a text density feature; and

b3) converting the table data encompassed by each table bounding box to a markup language representation.

5. The method of Claim 4 wherein receiving a page description language representation of the document for providing a list of words in the document and position information for the words includes receiving a PDF representation of the document, and wherein converting the table data encompassed by each table bounding box to a markup language representation includes converting the table data encompassed by each table bounding box to a HTML representation.

6. The method of Claim 4 wherein the table bounding box includes a top edge and a bottom edge; and wherein the step of expanding the table bounding box based on a text density feature comprises

b2_1) for each line determining a text density measure;

b2_2) for each line determining a change of text density between the current line and the previous line;

b2_3) if the change in text density reaches a predetermined threshold, marking the current line with a text density tag;

b2_4) expanding the top edge of the table bounding box in a first direction to one of a line previously marked by a text density tag and a line with a single word cluster; and

b2_5) expanding the bottom edge of the table bounding box in a second direction to one of a line previously marked by a text density tag and a line with a single word cluster.

7. A computer-readable medium having stored thereon sequences of instructions, said sequences of instructions including instructions which, when executed by a processor, cause said processor to perform the steps of:

a) receiving a page description language representation of a document for providing a list of words in the document and position information for the words; and

b) automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature.

8. The computer-readable medium of Claim 7 further containing instructions which, when executed by said processor, would cause said processor to perform the steps of:

b1) dividing the document into one or more pages;

b2) dividing each page into a plurality of lines;

b3) for each line, clustering the words of the line into one or more word clusters;

b4) automatically identifying table data in the document based on the number of word clusters for each line and the alignment of the word clusters between lines.

9. The computer-readable medium of Claim 8 further containing instructions which, when executed by said processor, would cause said processor to perform the steps of:

b4_1) using the word clusters to generate column position information; and

b4_2) updating the column position information by performing a union operation between the column position information of the previous line and the column position information of the current line.

10. The computer-readable medium of Claim 7 further containing instructions which, when executed by said processor, would cause said processor to perform the steps of:

b1) automatically determining a table bounding box for each table in the document;

b2) expanding each table bounding box based on a text density feature; and

b3) converting the table data encompassed by each table bounding box to a markup language representation.

11. The computer-readable medium of Claim 8 wherein, the computer-readable medium further containing instructions which, when executed by said processor, would cause said processor to perform the steps of:

b2_1) for each line determining a text density measure;

b2_2) for each line determining a change of text density between the current line and the previous line;

b2_3) if the change in text density reaches a predetermined threshold, marking the current line with a text density tag;

b2_4) expanding the top edge of the table bounding box in a first direction to one of a line previously marked by a text density tag and a line with a single word cluster; and

b2_5) expanding the bottom edge of the table bounding box in a second direction to one of a line previously marked by a text density tag and a line with a single word cluster.

12. A document processing system comprising:

a) a processor for executing programs; and

b) a table identification program for receiving a page description language representation of a document, the page description language representation providing a list of words in the document and position information for the words, and for automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature.

13. The document processing system of claim 12 wherein the table identification program further comprises:

b1) a bounding box generation module for receiving the list or words and for automatically generating a table bounding box for each table in the document based on the number of word clusters in each line.

14. The document processing system of claim 13 wherein the table identification program further comprises:

b2) a expansion module coupled to the bounding box generation module for receiving the table bounding box for each table in the document, wherein each table bounding box has a first edge and a second edge; the expansion module for expanding the first edge in a first direction to one of a line that has a single word cluster and a line that has been previously marked with a text density tag and for expanding the second edge in a second direction to one of a line that has a single

word cluster and a line that has been previously marked with a text density tag.

15. The document processing system of claim 13 wherein the table identification program further comprises:

b3) a conversion module coupled to the bounding box generation module for receiving the table bounding box for each table in the document, and for converting the words encompassed by the table bounding box into a markup language representation that maintains the table structure of each table.

16. The method of Claim 1 wherein the step of automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature further comprises:

b1) automatically identifying table data in the document based on one or more table headings.

17. The method of Claim 1 wherein the step of automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature further comprises:

b1) automatically identifying table data in the document based on one or more horizontal lines and vertical lines that separate rows or columns of the table.